Entropy measures and predictive recognition as mirrored in gating and lexical decision over multimorphemic Hungarian noun forms

Csaba Pléh together with Kornél Németh, Dániel Varga, Judit Fazekas, and Klára Várhelyi

Department of Cognitive Science
and MOKK Budapest University of Technology and Economics

Talk at the workshop
Quantitative measures in morphology
and morphological development

UCSD Center for Human Development, January 15-16, 2011

The team
Csaba Kornél Dani Klára Judit

Outline

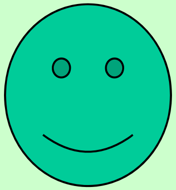
- The relevance of information theory for word processing
- The structure of Hungarian nouns and entropy
- Gating studies on word stem and information value
- Scrambled words and the onset superiority
- Effects of morphological complexity and information value on lexical decisions

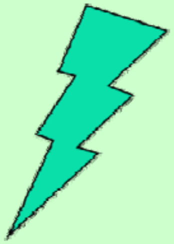
Stages in the relevance of Info Theory for language

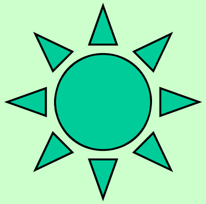
- Early enthusiasm statistics

Severe critic

New possibilities:







- 1950 Shannon
- Miller

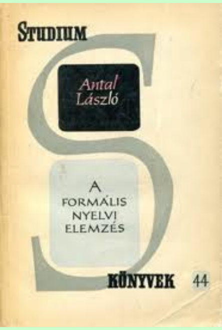
1960 Chomsky

G. Miller

1990 Kostic

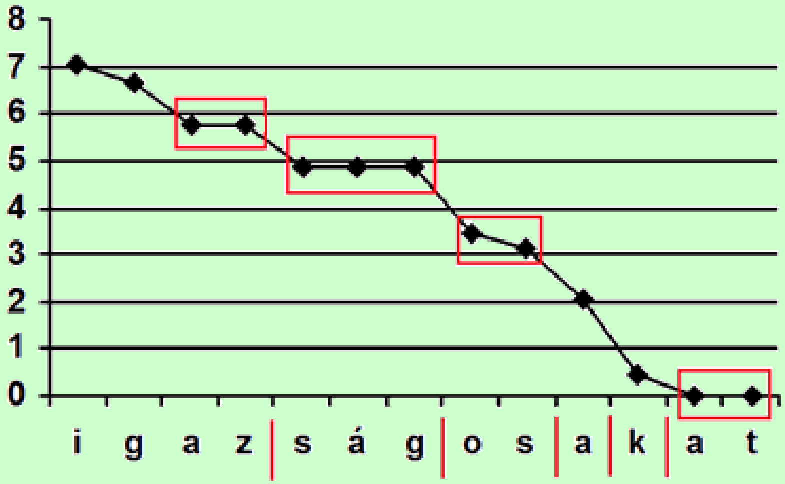
Saffran

Early proposals for info theory related morphology



- Antal László (1964) over words, the tendency is gradually decreasing entropy. Morpheme boundaries correspond to sudden drops in entropy.

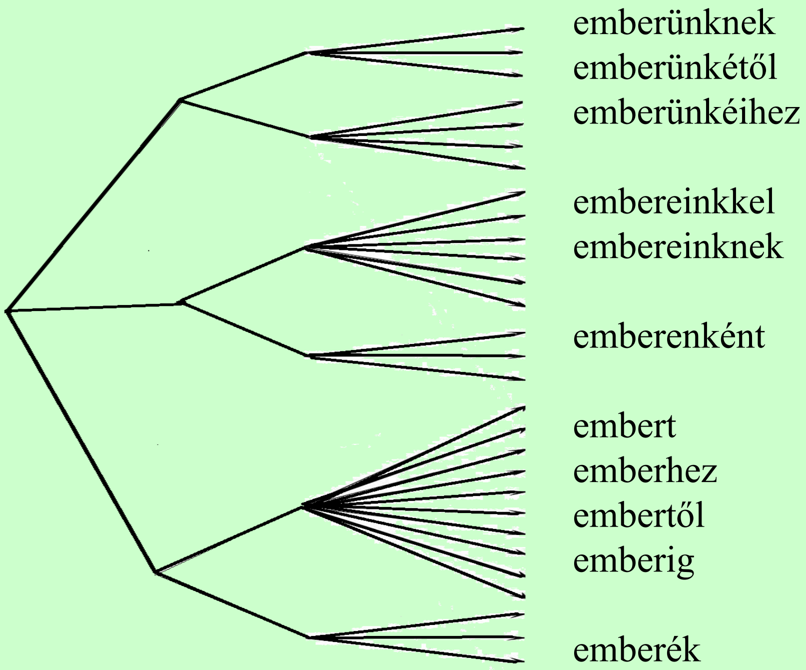
- *igaz-ság-os-ak-at* 'true-th-full-Plur-Accus' truesfulls



Morphemes are identified by plateaus of entropy.

Structure of Hungarian nouns

- stem – derivation – possession- possPlu- PossPers- Possessed- Plural-case



It is easy to assign numbers for these arborizations:

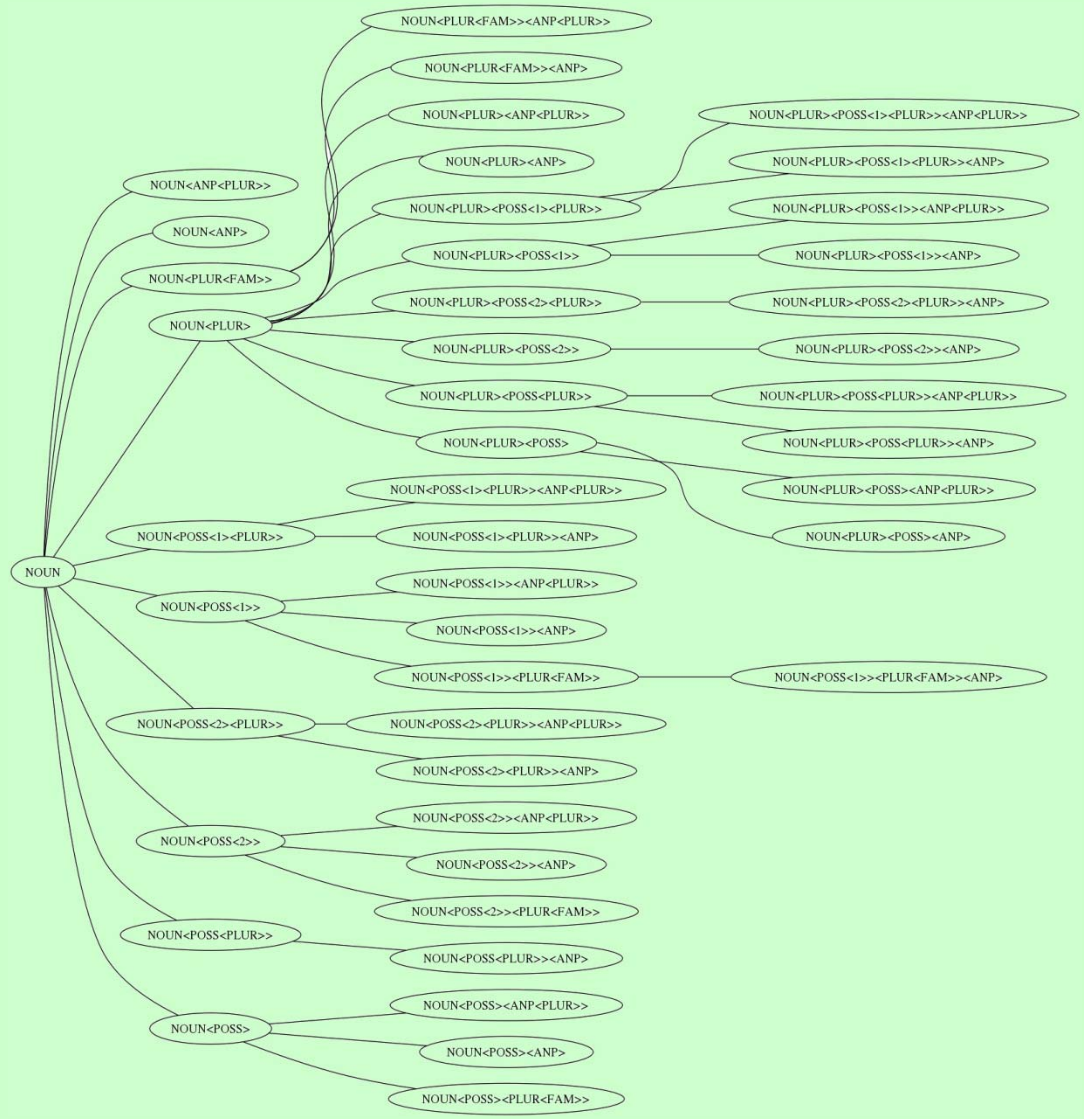
number of branches

token and type entropies

at different decision points

Structure of Hungarian nouns

Illustration: a small subtree of the tree of possible Hungarian noun inflections. The full tree is about 50 times larger.



Inflectional suffixes

ház *house*

ház-at *house ACC*

ház-ak *housePlur*

ház-am *house-PossMe*

ház-ak-at *housePlurAcc*

ház-a-m-at *housePoss1stSingAcc*

Derivational suffixes

ház *house*

ház-as '*housey' i.e. married*

ház-as-ság '*house-y-ness' i.e. marriage*

Both derivation and inflection

- ház *house*
- ház-as-ság *marriage*
- ház-as-ság-a-i-m-ban *marriage PossPlur1stSing-IN*

  ‘in my marriages’

Issues of morphology processing

1. **Segmentation** : morfotactics, primacy
2. **Lexical access :** analytic, holistic, mixed
   acoustic and ortoghraphic access files
3. **Formal combinatorics:** arguments
4. **Semantic integration:** transparency issues
5. **Stem allomorphy:** ? Separate routes?

Series of studies

- Gating
- Scrambling reconstruction
- Lexical decision

Questions

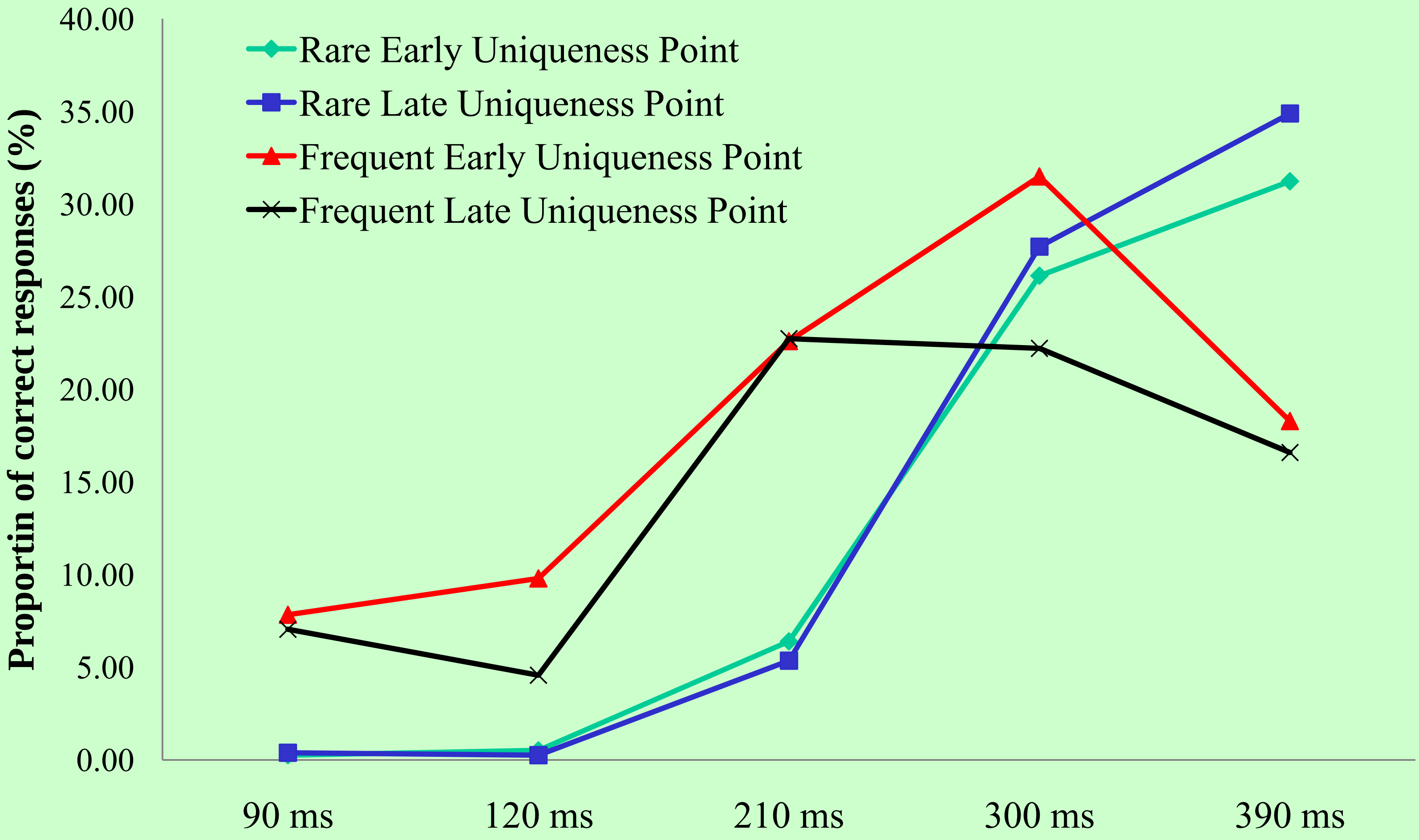
- Role of entropy related predictability in recognition
- The primacy issue in word recognition
- What happens with morphologically complex word forms?

Gating

- Four types of words

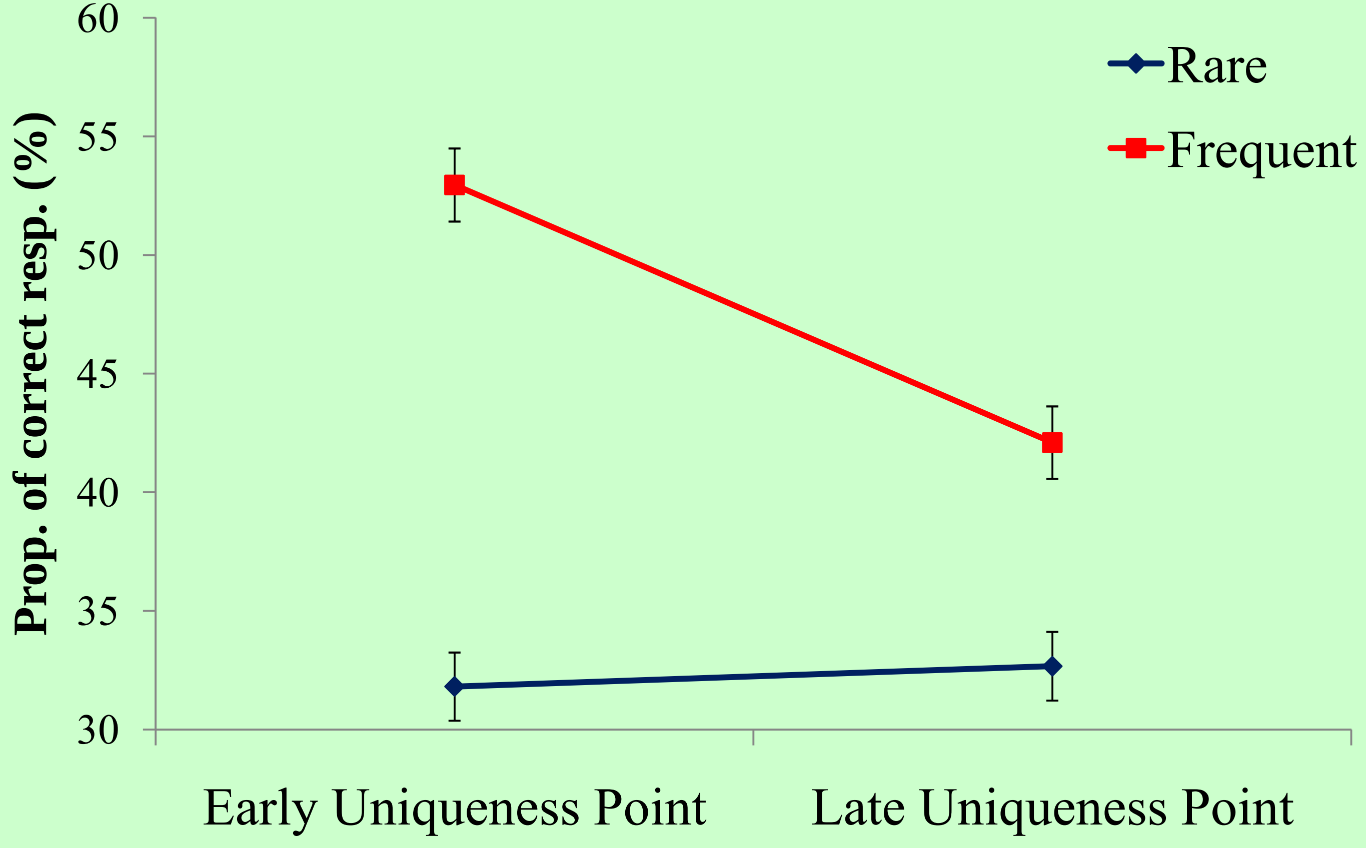
	Rare	Frequent
Early	*böllér* pigsticker	*kenyér* bread
Late	*pincsi* pekingese	*város* *town*

WEB based study

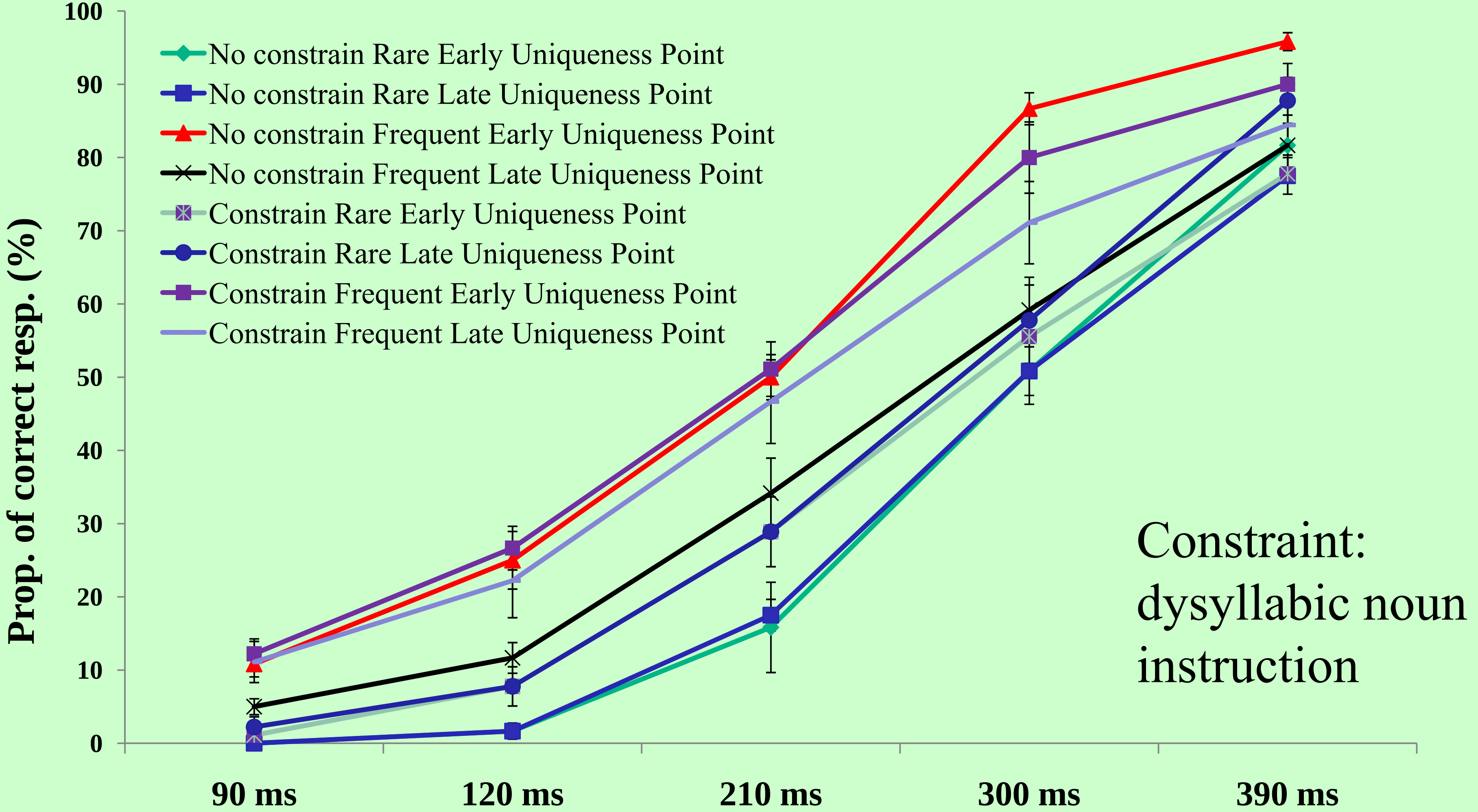


frequent words
were recognized in 300 msecs
rare words at the longest presentation

Lab gating study clear interaction between uniqueness point and frequency



Instructions also have a top down effect



Effects of constraints, frequency, and uniqueness points on gating recognition.

Summary of the gating effects

- a syllable length segment can lead to 50 percent correct recognition in line with prediction e.g. from cohort theory
- this is related to the uniqueness point issue as well, the more rivals to a word, the later the recognition point.
- frequency facilitates recognition
- in rare words there is a more strictly bottom up recognition process, competing neighbors have no effect in their case.
- top down effect of constraints in the instructions: grammatical and morphotactic constraints also played a role in Hungarian.

Comparing with the MOKK corpus based entropies

corpus	pages (million)	token (million)	type (million)
full	3,5	1486	19,1
60% Foreign excluded	3,125	1310	15,4
92% Only text with diacritics	1,918	928	10,9
96% Typos as in normal text	1,221	589	7,2

Effects of entropy: different indicators

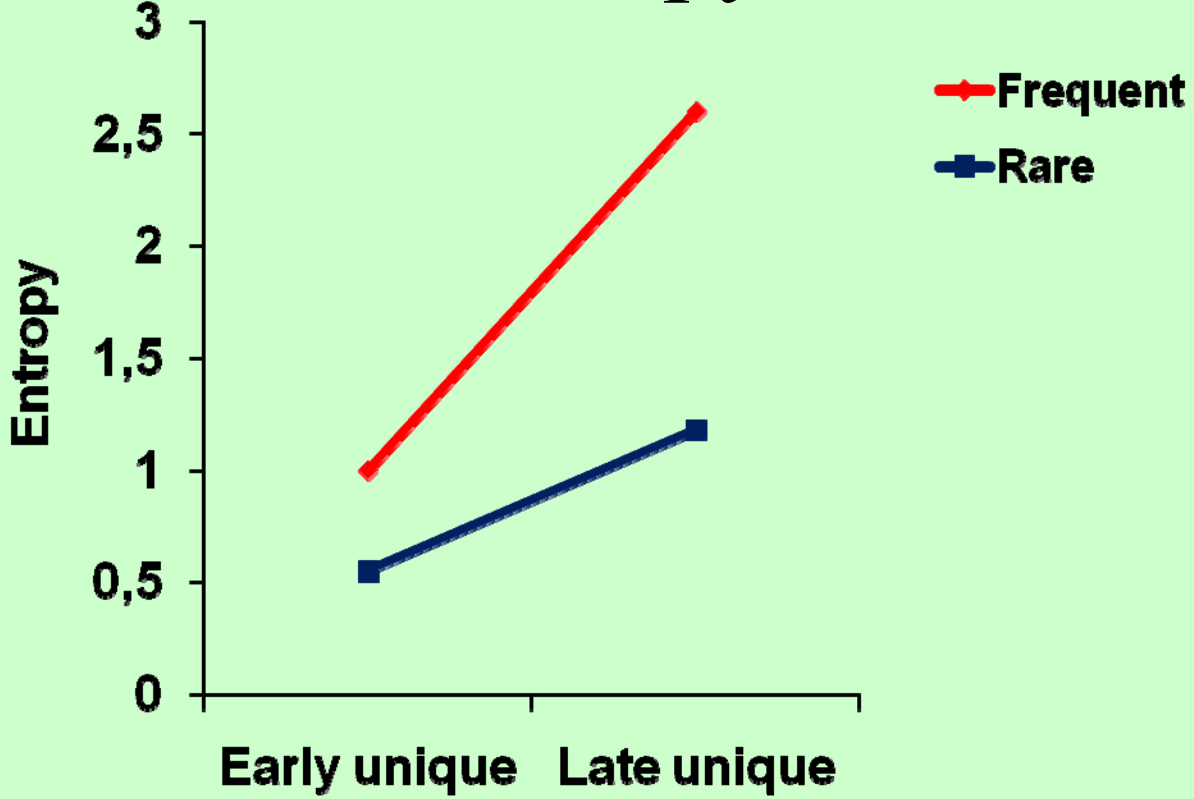
- **prefixtypeoccurenceslog** number of word forms in the corpus starting with the given prefix. We work with the base 2 logarithm of this value.

  prefixfreqlog –Number of tokens in our corpus starting with the given prefix. (logarithm)
  entropy – Entropy of the corpus, conditioned on the given prefix. Informally, it is our amount of uncertainty about an unknown word from the corpus, when we are told its prefix. Formally, it is defined as

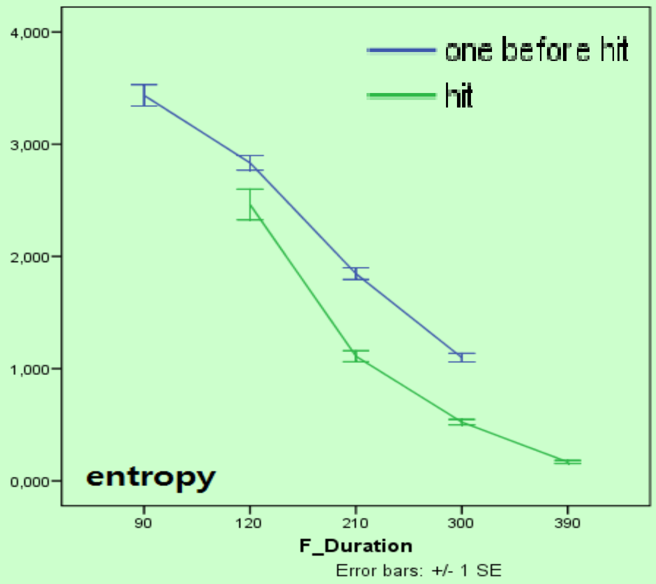
$$\mathrm{H}(\mathrm{W} \mid \mathrm{x}) = \sum_{\mathrm{w} \in \mathrm{W}} p(\mathrm{w} \mid \mathrm{x}) \log_2 p(\mathrm{w} \mid \mathrm{x})$$

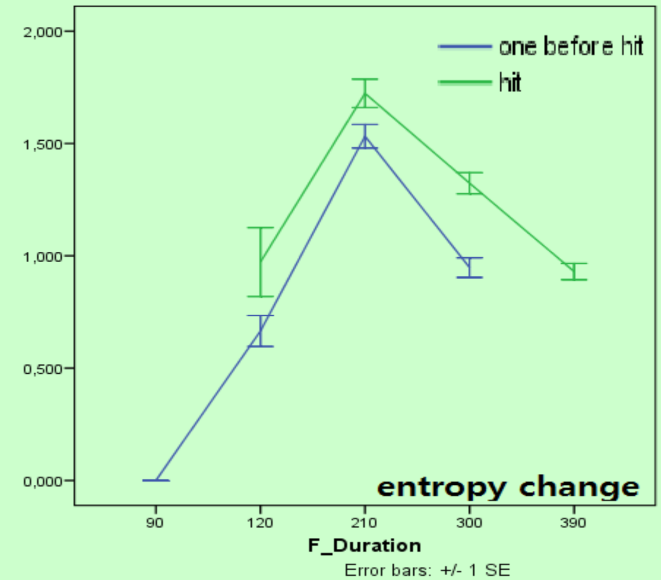
- **entropychange** – The decrease in entropy when compared to the previous gate.

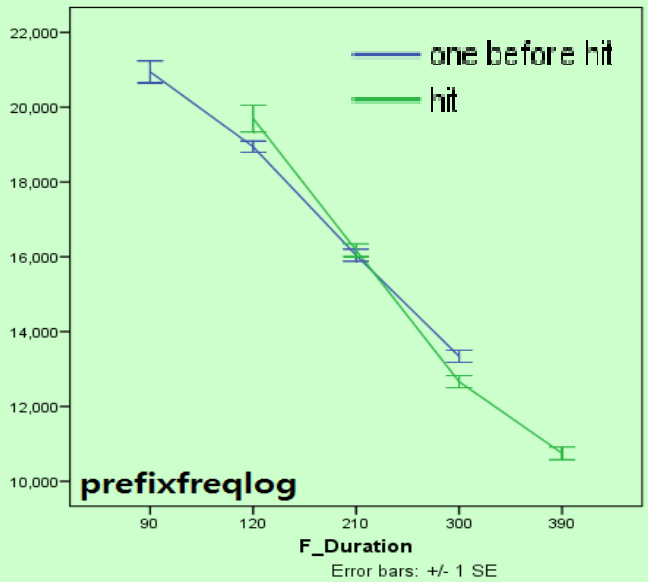
Entropy is greater in frequent words at point 4
Uniquness point: decrease in entropy

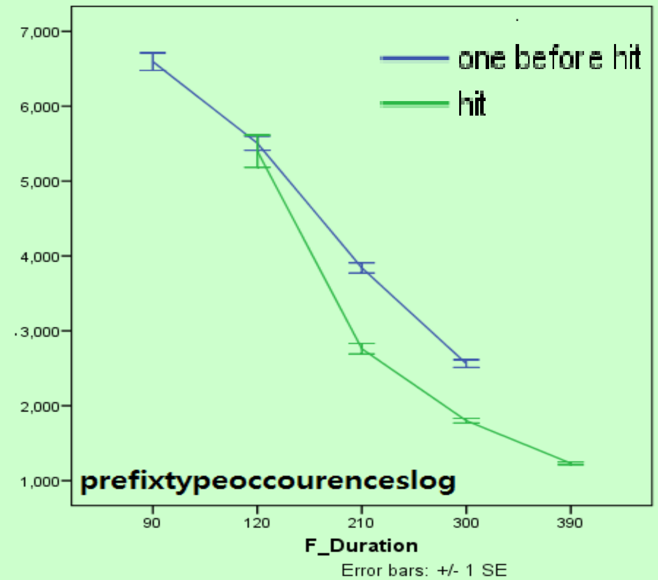


Gating gates for the different entropy and word competition effects



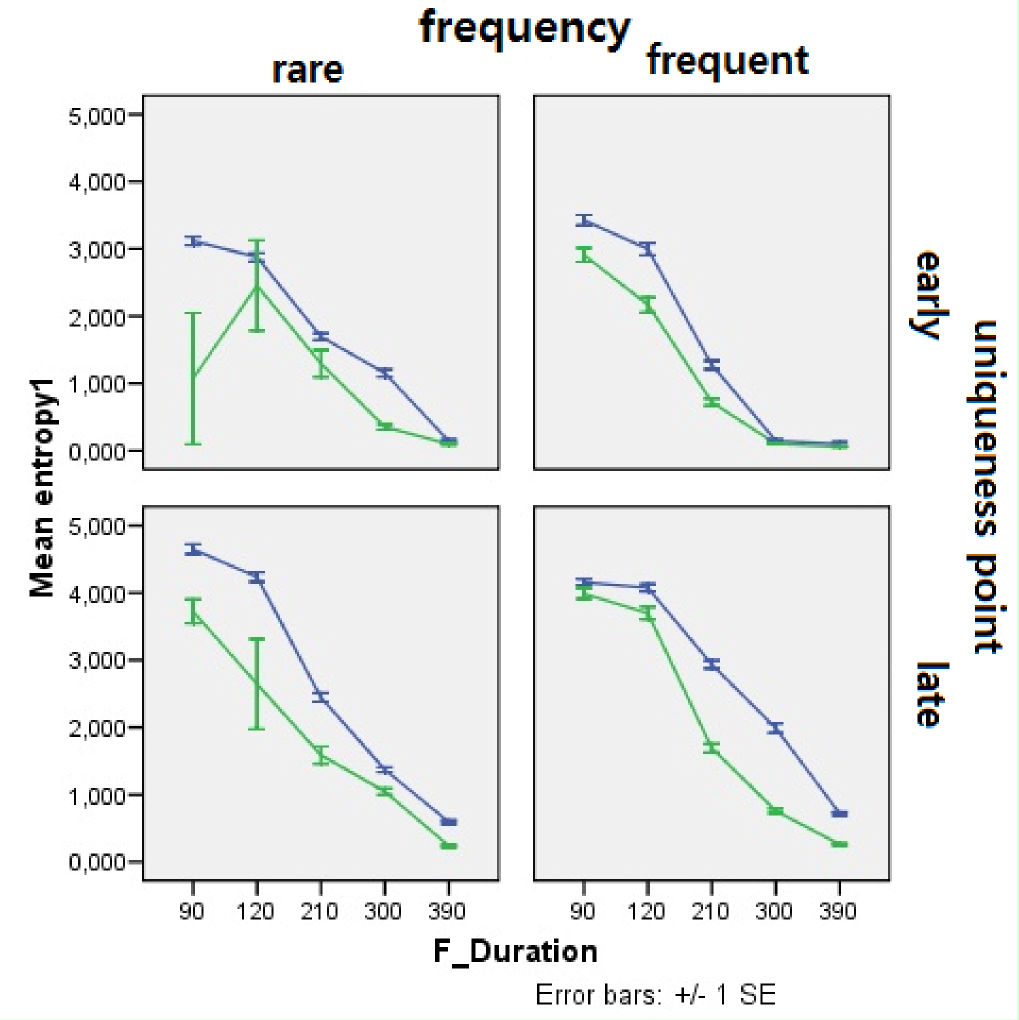






Entropy and uniqueness point

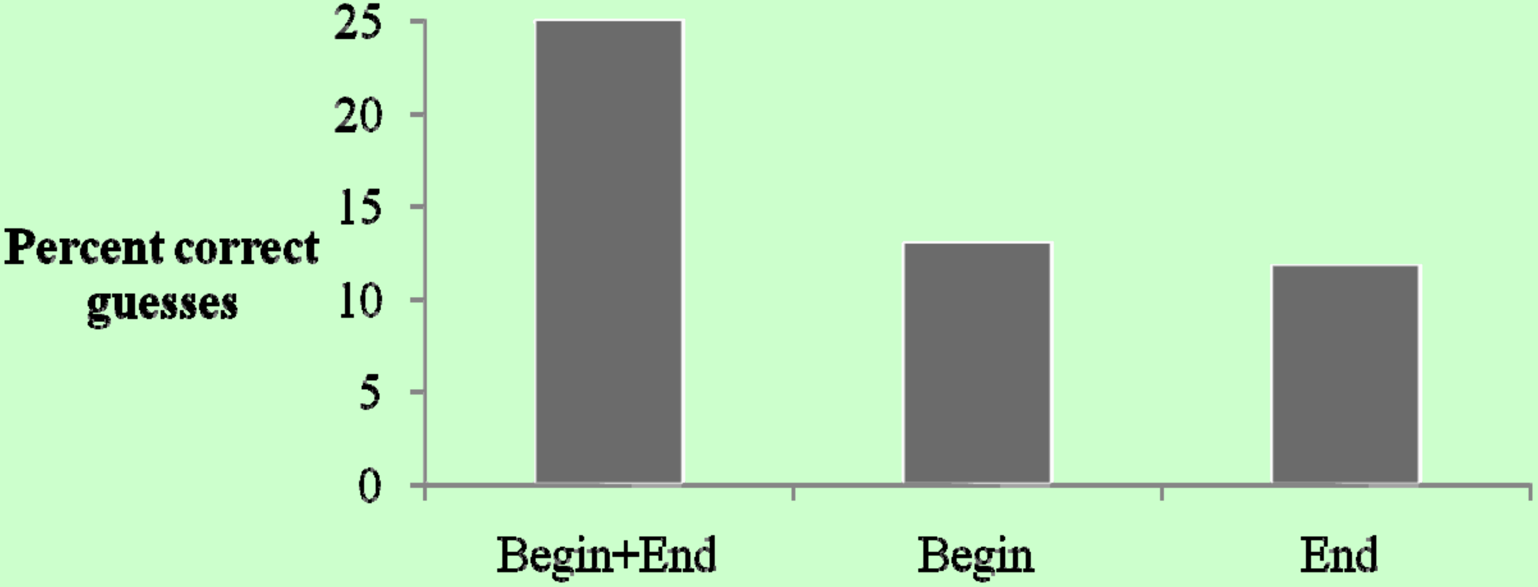
Entropy has a statistically significant relationship with recognition rate, even when we control for uniqueness point and frequency.



Entropy effects

- All measures have a significant effect on recognition rate.
- The effect of entropychange (a highly non-monotonous function of prefix-length) means that the recognition point follows a sudden drop of the entropy value, which is the hypothesis we started from.
- The effect of entropy when controlled for uniqueness point can be interpreted as showing that entropy is a refinement of the naive uniqueness point metric.

The importance of beginnings: reconstruction of scrambled words

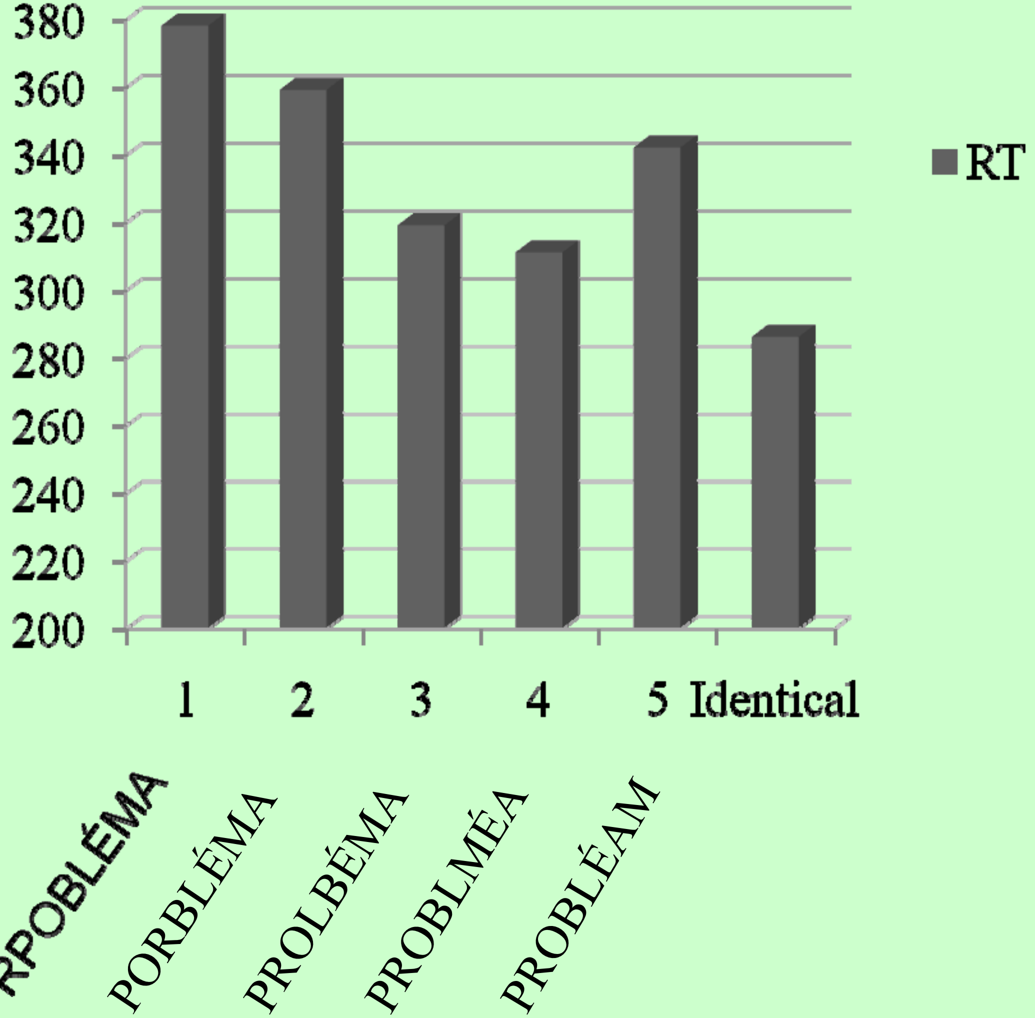


szitimkus **sötbténe** **mléény**

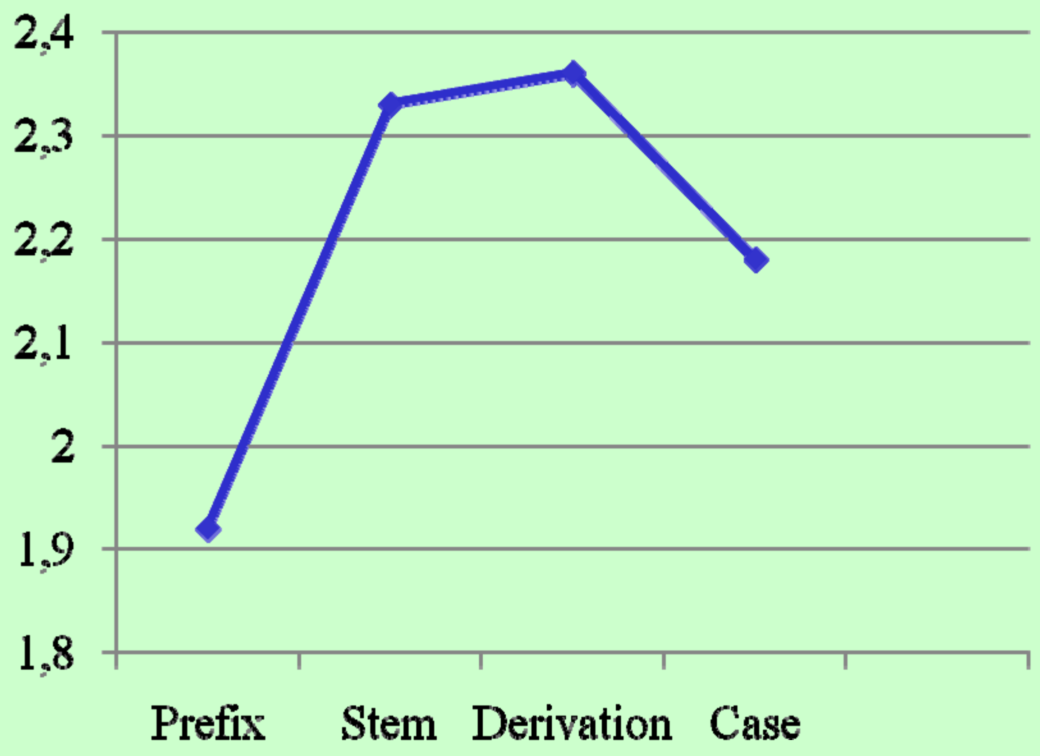
The priming study

RPROBLÉMA

Priming effect of beginnings : spoiled word slows down at beginning and end



Decisions over long multimophemic words: Pléh and Juhász 1995



bathtub effect
Aithchison

Systematic study

- stem *böllér* 'sticker'
- plural *böllér-ek* 'stickers'
- case *böllér-nek* 'stickerDAT'
- Plur + case *böllér-ek-nek* 'stickersDAT'

Böll**á**r

Böllér-**a**k böllér**u**k

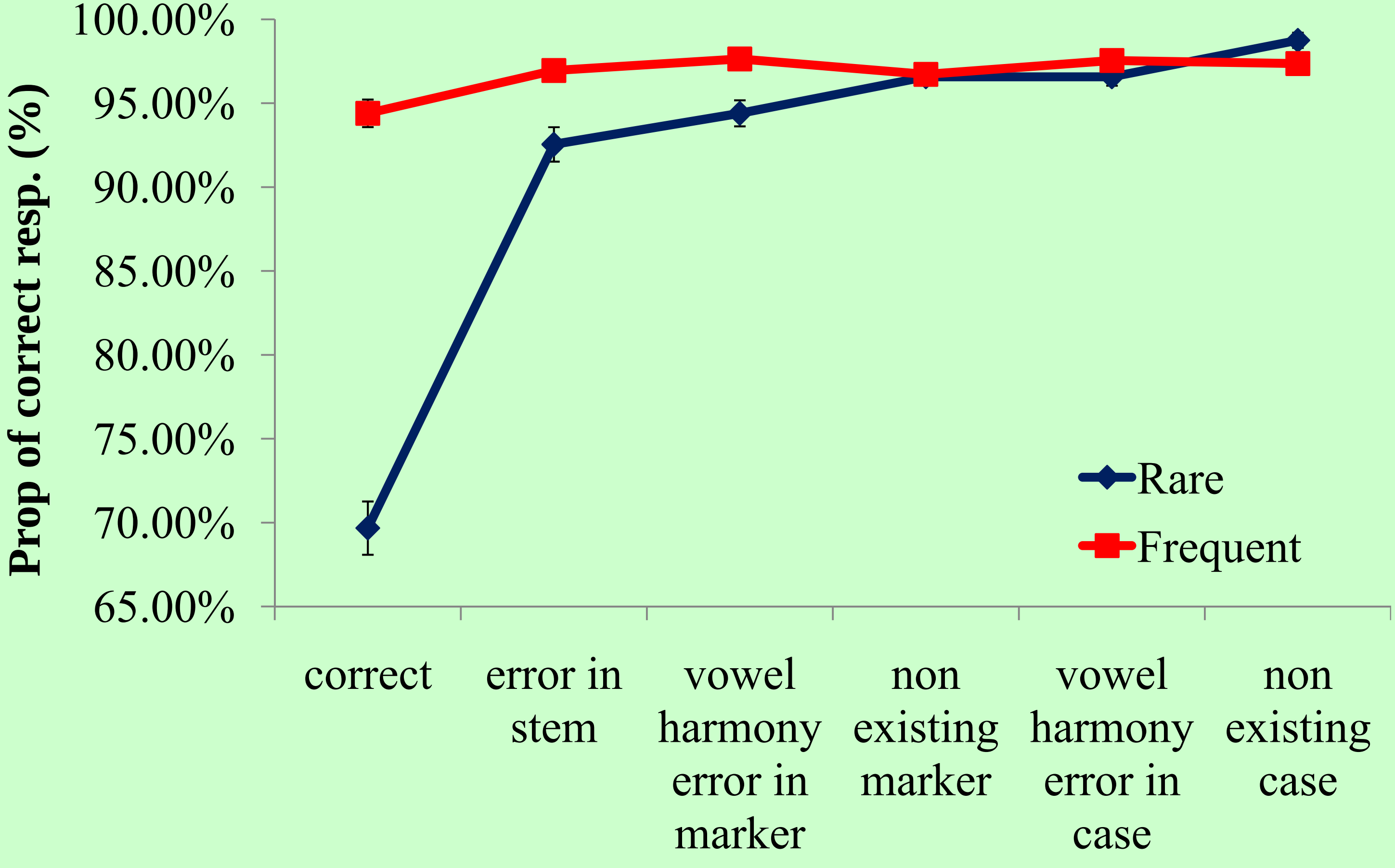
Böllér-n**a**k böllér-n**u**k

Böllér-**a**k-nek böllér-**u**k-nek

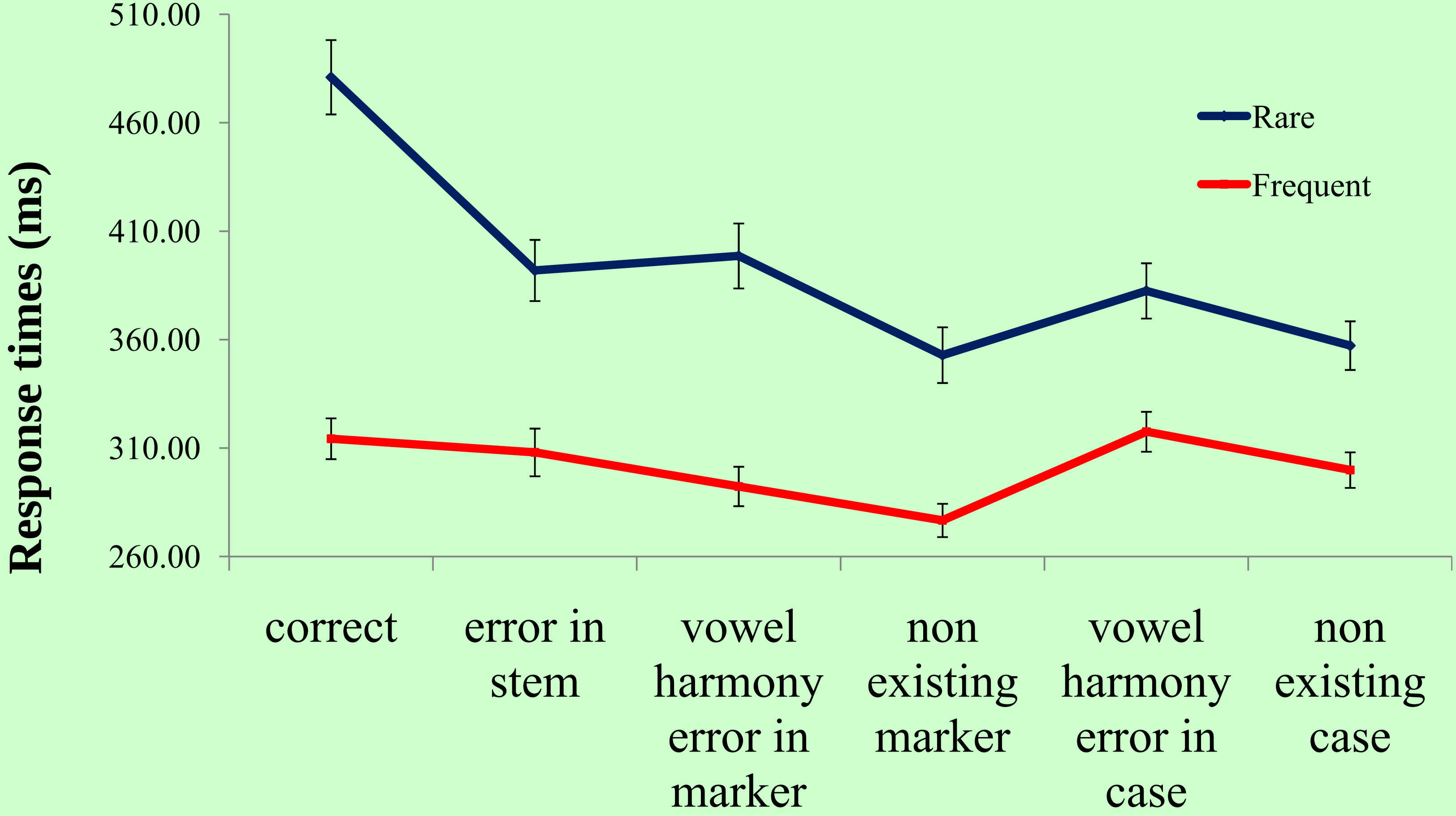
böllérnek

bölléruk

Correctness of acceptance-rejection as a function of word structure



Reaction time data in the lexical decision task



Paired comparions

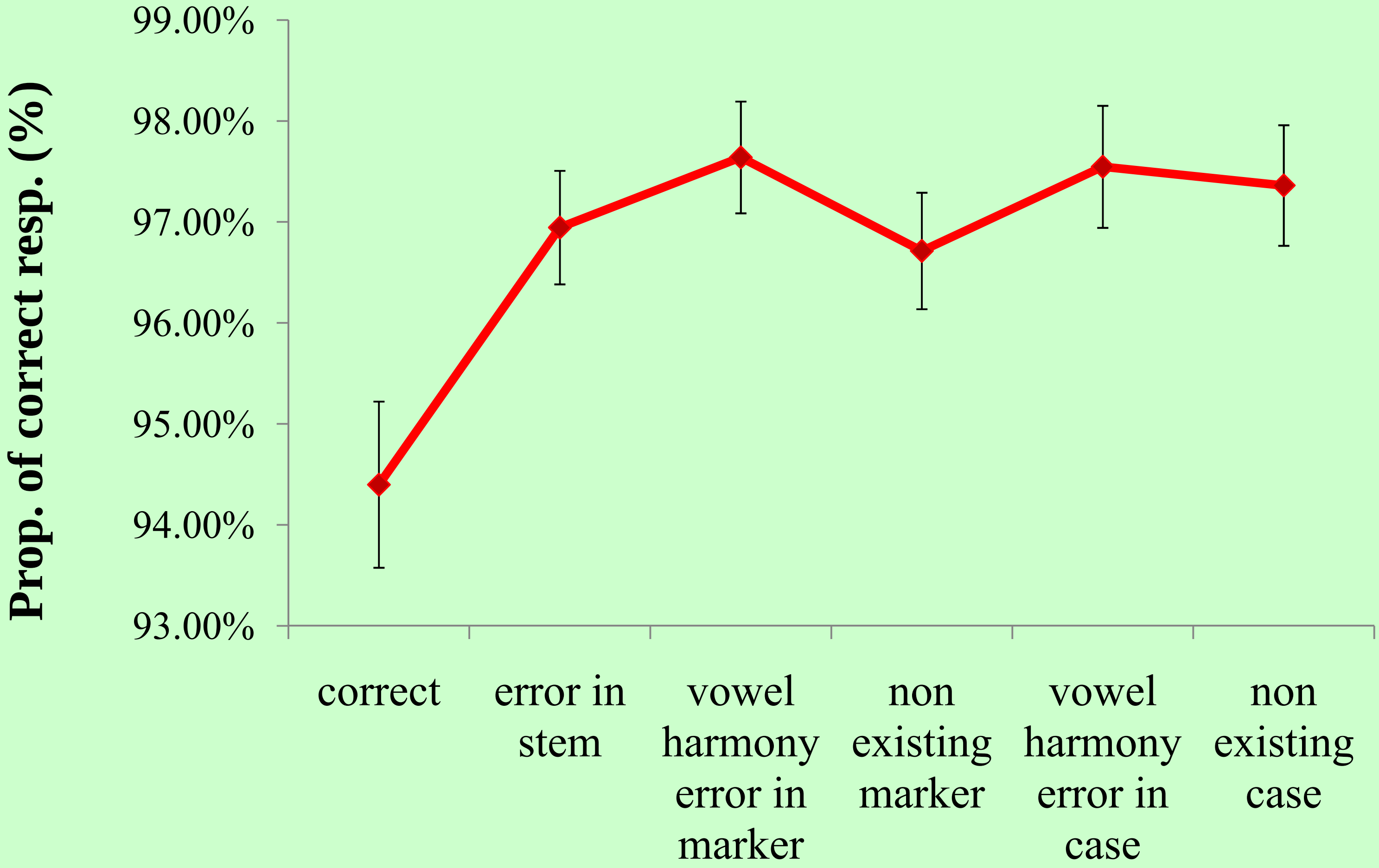
Correct forms were slower to be accepted.

Non existing stems were slower to be rejected than non existing markers (i.e. word middle errors) or case markers.

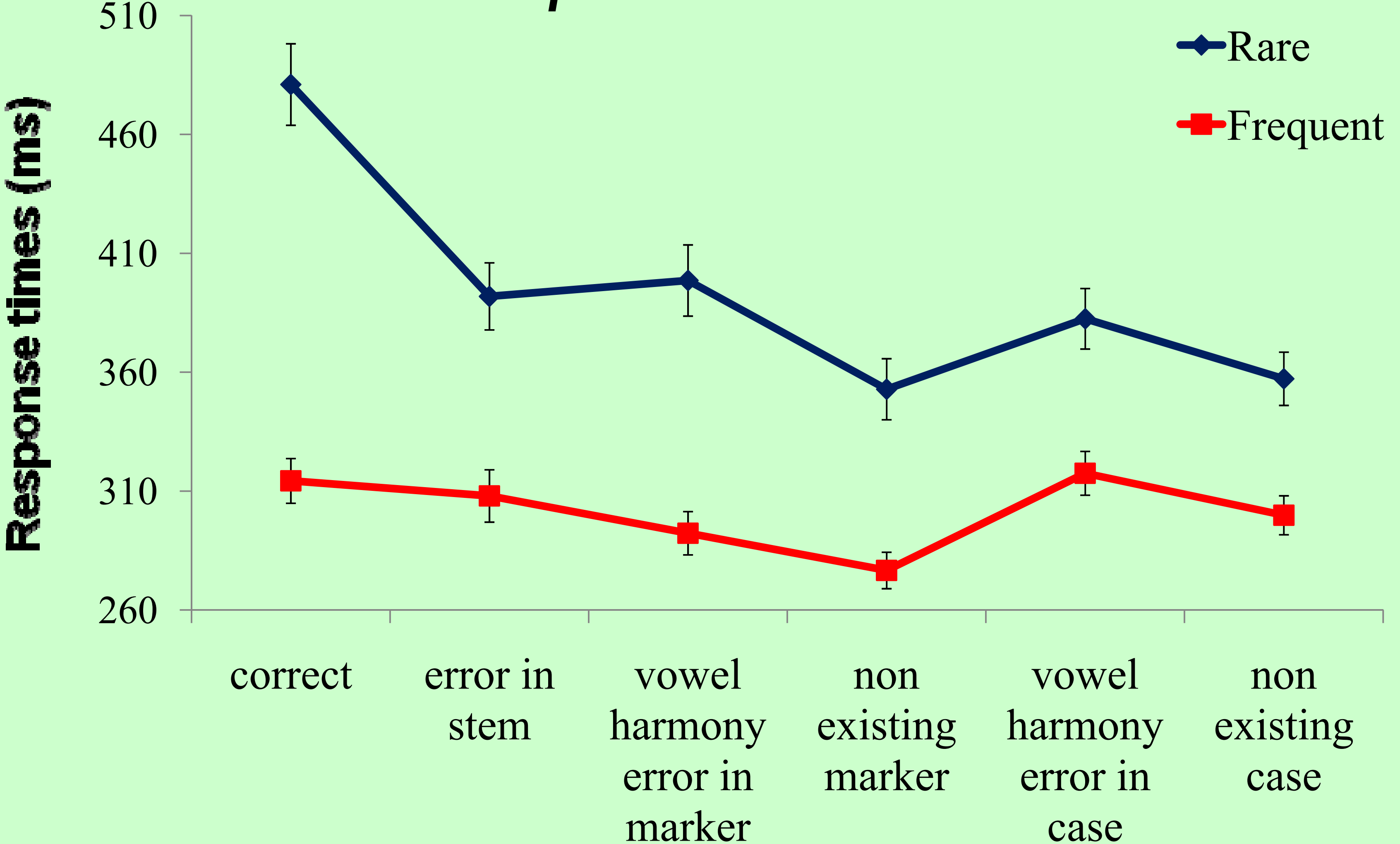
No clear bath tub effect in the reaction times.

Word ending case marker errors were recognized slower.

In frequent items word middle distortions lower performance



Reaction times in both frequent and rare items are fast in word middle non-existing form errors



Planned further analysis

- Correlate decisions and times with rivaling tokens at the manipulated point
- Correlate with different entropy measures
- Combined entropy of the stems and the endings

Summary

- There is a strong word onset primacy in Hungarian as well
- Word recognition is more sensitive to entropy values and morphological structure than to frequency itself
- Entropy change is important in explaining neighborhood effects